Research Data Management & Data Management Plans (DMPs)

Sarah Jones
DCC, University of Glasgow
sarah.jones@glasgow.ac.uk
Twitter: @sjDCC
What is research data management?

“The active management and appraisal of data over the lifecycle of scholarly and scientific interest”

Data management is part of good research practice
Why manage research data?

• To make research easier!
• To stop yourself drowning in irrelevant stuff
• In case you need the data later
• To avoid accusations of fraud or bad science
• To share data so others can use and learn from it
• To get credit for producing the data
• Because somebody else said to do so
RCUK Common Principles in brief

1. Make data openly available where possible
2. Have policies & plans. Preserve data of long-term value
3. Metadata for discovery / reuse. Link to data from publications
4. Be mindful of legal, ethical and commercial constraints
5. Allow limited embargoes to protect the effort of creators
6. Acknowledge sources to recognise IP and abide by T&Cs
7. Ensure cost-effective use of public funds for RDM

www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
Ultimately funders expect:

- **timely release of data**
  - once patents are filed or on (acceptance for) publication

- **open data sharing**
  - minimal or no restrictions if possible

- **preservation of data**
  - typically 5-10+ years if of long-term value

See the RCUK Common Principles on Data Policy:
[www.rcuk.ac.uk/research/Pages/DataPolicy.aspx](http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx)
Why make data available?

“It was *never* acceptable to publish papers without making data available.”

- Ewan Birney

#OpenData
#OpenScience

Original image via doi:10.1038/461145a. “Research cannot flourish if data are not preserved and made accessible. Data management should be woven into every course in science.” - Nature 461, 145
Sharing leads to breakthroughs

"It was unbelievable. It's not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr. John Trojanowski, University of Pennsylvania

...and increases the speed of discovery
What is involved in RDM

- Data Management Planning
- Data creation
- Annotating / documenting data
- Analysis, use, versioning
- Storage and backup
- Publishing papers and data
- Preparing for deposit
- Archiving and sharing
- Licensing
- Citing...
If researchers plan to share data....

- Have they got consent for sharing?
- Do licences / agreements permit sharing?
- Is the data in suitable formats?

Decisions made early on affect what can be done later
Some formats are better for long-term

It’s preferable to opt for formats that are:
- Uncompressed
- Non-proprietary
- Open, documented
- Standard representation (ASCII, Unicode)

Data centres may have preferred formats for deposit e.g.

<table>
<thead>
<tr>
<th>Type</th>
<th>Recommended</th>
<th>Non-preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabular data</td>
<td>CSV, TSV, SPSS portable</td>
<td>Excel</td>
</tr>
<tr>
<td>Text</td>
<td>Plain text, HTML, RTF PDF/A only if layout matters</td>
<td>Word</td>
</tr>
<tr>
<td>Media</td>
<td>Container: MP4, Ogg Codec: Theora, Dirac, FLAC</td>
<td>Quicktime H264</td>
</tr>
<tr>
<td>Images</td>
<td>TIFF, JPEG2000, PNG</td>
<td>GIF, JPG</td>
</tr>
<tr>
<td>Structured data</td>
<td>XML, RDF</td>
<td>RDBMS</td>
</tr>
</tbody>
</table>

Further examples: [http://www.data-archive.ac.uk/create-manage/format/formats-table](http://www.data-archive.ac.uk/create-manage/format/formats-table)
Documentation and metadata

**Metadata**: basic info e.g. title, author, dates, access rights...

**Documentation**: context, workflows, methods, code, data dictionary...

Use standards wherever possible for interoperability

[www.dcc.ac.uk/resources/metadata-standards](http://www.dcc.ac.uk/resources/metadata-standards)
Tools for managing data

Managing Active Research Data

Resources to help researchers manage data from its point of creation, facilitating its productive use in the present, but also establishing the support structures necessary to ensure its future survival.

Active Data Storage (2)
Flexible storage for evolving data.

Data Management Planning (2)
Help in meeting funder requirements for data management plans.

Persistent ID Assignment (4)
Creating unique identifiers for digital objects.

Workflow and Lab Notebook Management (5)
Digitising the traditional lab notebook and automating scientific workflows.

www.dcc.ac.uk/resources/external/tools-services/managing-active-research-data
Where to store data?

• Your own device (PC, flash drive, etc.)
  – And if you lose it? Or it breaks?

• Departmental drive or university filestore
  – Should be more robust with automated back-up

• “Cloud” storage
  – Do they care as much about your data as you do?
How to backup?

• 3... 2... 1... backup!
  – at least 3 copies of a file
  – on at least 2 different media
  – with at least 1 offsite

• Use managed services where possible e.g. University filestores rather than local or external hard drives

• Ask central or local IT team for advice
Licensing data for reuse

Outlines pros and cons of each approach and gives practical advice on how to implement data licences

CREATIVE COMMONS LIMITATIONS

- **NC** Non-Commercial
  - What counts as commercial?

- **SA** Share Alike
  - Reduces interoperability

- **ND** No Derivatives
  - Severely restricts use

[www.dcc.ac.uk/resources/how-guides/license-research-data](http://www.dcc.ac.uk/resources/how-guides/license-research-data)
Archiving: data repositories

http://service.re3data.org/search

Zenodo
- OpenAIRE-CERN joint effort
- Multidisciplinary repository
- Multiple data types
  - Publications
  - Long tail of research data
- Citable data (DOI)
- Links to funder, publications, data & software

www.zenodo.org
Managing and sharing data: a best practice guide

http://data-archive.ac.uk/media/2894/managingsharing.pdf
What is a data management plan?

A brief plan written at the start of your project to define:

• how your data will be created?
• how it will be documented?
• who will access it?
• where it will be stored?
• who will back it up?
• whether (and how) it will be shared & preserved?

DMPs are often submitted as part of grant applications, but are useful whenever researchers are creating data.
Which UK funders require a DMP?

<table>
<thead>
<tr>
<th>Research Funders</th>
<th>Policy Coverage</th>
<th>Policy Stipulations</th>
<th>Support Provided</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Published outputs</td>
<td>Data</td>
<td>Time limits</td>
</tr>
<tr>
<td>AHRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>BBSRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>CRUK</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>EPSRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>ESRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>MRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>NERC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>STFC</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

DCC Checklist for a DMP

• 13 questions on what’s asked across the board

• Prompts / pointers to help researchers get started

• Guidance on how to answer

Common themes in DMPs

1. Description of data to be collected / created
   (i.e. content, type, format, volume...)

2. Standards / methodologies for data collection & management

3. Ethics and Intellectual Property
   (highlight any restrictions on data sharing e.g. embargoes, confidentiality)

4. Plans for data sharing and access
   (i.e. how, when, to whom)

5. Strategy for long-term preservation
A useful framework to get started

Framework for Creating a Data Management Plan

This framework can be used as an outline in assembling data management plans to accompany grant applications. Note that some funders have page limits for data management plans—NSF limits plans to two pages.

Elements of a Data Management Plan

This list of elements is informed by a gap analysis that ICPSR conducted of existing recommendations for data management plans and other forms of guidance made available for researchers generating data. The result of the gap analysis was a comparison of existing forms of guidance. Elements that are highly recommended for inclusion in effective data management plans are noted.

See our bibliography for additional readings germane to the elements of a data management plan.

Data Description (Recommended)

Provide a brief description of the information to be gathered -- the nature, scope, and scale of the data that will be generated or collected.

Why this is important

A good description of the data to be collected will help reviewers understand the characteristics of the data, their relationship to existing data.

Example 1:

This project will produce public-use nationally representative survey data for the United States covering Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life.

Example 2:

This project will generate data designed to study the prevalence and correlates of DSM III-R psychiatric disorders and patterns and correlates nationally representative sample of over 8000 respondents. The sensitive nature of these data will require that the data be released through a confidentiality agreement.

Think about why the questions are being asked – why is it useful to consider that topic?

Look at examples to help you understand what to write

www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html
Help from the DCC

A web-based tool to help researchers write data management plans

https://dmponline.dcc.ac.uk

www.dcc.ac.uk/resources/how-guides/develop-data-plan
Tips to share on writing DMPs

• Keep it simple, short and specific
• Seek advice - consult and collaborate
• Base plans on available skills and support
• Make sure implementation is feasible
• Justify any resources or restrictions needed

www.youtube.com/watch?v=7OJtiA53-Fk
Who should pay for RDM?

Funding Research Data Management
"A conversation with the funders"

The DCC held a special event on this topic in the UK, but there’s still a long way to go

[Website Link]

www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf-special-event-funding-research-data-management
How should costs be recovered?

It is for institutions to decide which elements to recover:
• Directly (i.e. directly allocated or incurred on grants)
• Indirectly (by using QR funding or indirects on grants)

If the cost of RDM services and infrastructure is already recovered via your institution’s indirect cost rate, it cannot also be applied for as a direct cost.

Guiding principle: don’t charge for the same thing twice
What to charge and how?

Direct costs

• In-project costs that must be incurred before the end of the grant
• Potentially hardware, staff, expenses, costs of preparing data for deposit...
• Could include charges levied by repositories (pay once store forever)
• May also include costs ordinarily recovered indirectly (e.g. storage) if the requirement is exceptional and exceeds the norm

Indirect costs

• The general cost of providing RDM services and infrastructure
• Designated data services should be used if provided
• Outsourcing to a third-party is also an option

Remember to make a clear justification for any costs
Be specific for each grant

• A flat rate charging structure for RDM services (e.g. 10% of each grant) is not appropriate.

• The value of a research grant is not a good proxy for the volume or complexity of data it may generate.

• Base costs on each specific case and make a clear justification for them. They should be auditable.
Key messages

• All costs are eligible, but:
  – Direct costs must be incurred before a grant ends
  – Nothing can be double funded (recovered indirectly and as a direct)

• Researchers are expected to use designated data repositories.

• There is no rule of thumb to measure the proportion of a grant that may acceptably be spent on research data management.

• A clear justification of resources is needed for each specific case.
Thanks – any questions?

DCC guidance, tools and case studies: www.dcc.ac.uk/resources

Follow us on twitter: @digitalcuration and #ukdcc
How to support researchers with DMPs

What role do (or should) you play?
- Raising awareness of requirements?
- Providing basic advice?
- Pointing researchers to relevant services?
- Helping to write plans and cost in RDM support?

What do you need to support researchers?
- Guidance on the intranet / FAQs?
- Training (more than today?)
- More time/resource?